Kernels and Submodels of Deep Belief Networks

Guido F. Montúfar
Department of Mathematics
Pennsylvania State University
University Park, PA 16802
gfm10@psu.edu

Jason Morton
Department of Mathematics
Pennsylvania State University
University Park, PA 16802
morton@math.psu.edu

Abstract

We study the mixtures of factorizing probability distributions represented as visible marginal distributions in stochastic layered networks. We take the perspective of kernel transitions of distributions, which gives a unified picture of distributed representations arising from Deep Belief Networks (DBN) and other networks without lateral connections. We describe combinatorial and geometric properties of the set of kernels and products of kernels realizable by DBNs as the network parameters vary. We describe explicit classes of probability distributions, including exponential families, that can be learned by DBNs. We use these submodels to bound the maximal and the expected Kullback-Leibler approximation errors of DBNs from above depending on the number of hidden layers and units that they contain.

1 Introduction

Deep belief networks (DBNs) are a kind of learning machine introduced originally in [10]. They are used to extract *features* from data, often by an unsupervised pretraining step, so their properties as generative models and their expressive power are also of interest, see [2, 23, 11, 15]. A DBN can be seen as a concatenation of modules that implement kernel transitions (stochastic linear maps) of probability vectors. We describe this perspective in Section 2, and the geometry and combinatorics of the set of kernels that DBNs can represent, in Section 3. See Figure 1.

The deep belief network probability model $\mathrm{DBN}(n_0, n_1, \dots, n_l)$ with layers of widths $n_0, \dots, n_l$ is the set of marginals $P(h^0) = \sum_{h^1\in\{0,1\}^{n_1}} \cdots \sum_{h^l\in\{0,1\}^{n_l}} P(h^0, h^1, \dots, h^l)$ for all $h^0 \in \{0,1\}^{n_0}$, of all joint probability distributions on the states of a layered network. The top layer has bipartite undirected connections, with subsequent layers bipartite and downward-directed, giving joint unmarginalized probabilities:

$$P(h^0, h^1, \dots, h^l) = \Big(\prod_{k=1}^{l-1} P(h^{k-1}|h^k)\Big) P(h^{l-1}, h^l) \ , \tag{1}$$

for all $(h^0, \dots, h^l) \in \{0,1\}^{n_0} \times \cdots \times \{0,1\}^{n_l}$, where

$$P(h^{l-1}, h^l) = \frac{1}{Z} \exp\left(h^l B^l + h^l W^l h^{l-1} + B^{l-1} h^{l-1}\right) \ , \text{ and} \tag{2}$$

$$P(h^{k-1}|h^k) = \frac{1}{Z_{h^k}} \exp\left(h^k W^k h^{k-1} + B^{k-1} h^{k-1}\right) \ . \tag{3}$$

Here $h^k = (h^k_1, \dots, h^k_{n_k}) \in \{0,1\}^{n_k}$ denotes the states of the units in the kth layer; $W^k \in \mathbb{R}^{n_k \times n_{k-1}}$ is a matrix of connection weights between units from the kth and $(k-1)$th layer; $B^k \in \mathbb{R}^{n_k}$ is a vector of bias weights of the units in the kth layer; $Z = \sum_{h^{l-1},h^l} \exp(h^l W^l h^{l-1} +$

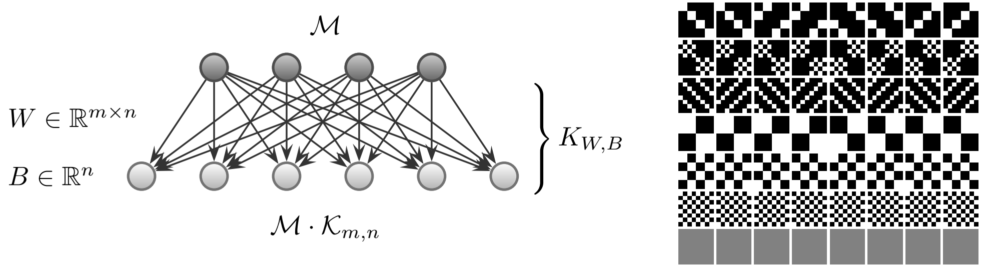
$\mathcal{M}$
$W \in \mathbb{R}^{m\times n}$
$B \in \mathbb{R}^n$
$K_{W,B}$
$\mathcal{M}\cdot\mathcal{K}_{m,n}$

Figure 1: Left: A network module that realizes stochastic transitions $\mathcal{K}_{m,n}$ from the set of distributions $\mathcal{M} \subseteq \Delta_{2^m-1}$ on the top layer, to probability distributions $\mathcal{M}\cdot\mathcal{K}_{m,n} \subseteq \Delta_{2^n-1}$ on the bottom layer, see eq (8). Right: The kernels $K_{W,B} = \mathrm{K}_p$ in $\mathcal{K}_{3,3} \subset \mathbb{R}^{8\times 8}$ described in Proposition 4.

$B^{l-1}h^{l-1} + b^l h^l)$ is a normalization constant that depends on W^l, B^{l-1}, B^l; and $Z_{h^{k+1}} = \sum_{h^k} \exp(h^{k+1}W^{k+1}h^k + B^k h^k)$ is a normalization constant that depends on W^{k+1}, B^k, and h^{k+1}. The total number of parameters of this model is $d = (\sum_{k=1}^{l} n_{k-1}n_k) + (\sum_{k=0}^{l} n_k)$, treating the layer widths $n_0,\dots,n_l$ as hyperparameters.

A *restricted Boltzmann machine* (RBM) [22, 6, 9] is formally the same as a DBN with only one hidden layer. The model $\mathrm{RBM}_{n,m} = \mathrm{DBN}(n,m)$ is the set of probability distributions on $\{0,1\}^n$ of the form $P(v) = \frac{1}{Z}\sum_{h\in\{0,1\}^m} \exp\big(hWv + Ch + Bv\big)$ for all $v \in \{0,1\}^n$.

We denote by Δ_{2^n-1} the simplex of probability distributions on $\{0,1\}^n$. Its vertices are the point measures δ_x, $x \in \{0,1\}^n$.

Sutskever and Hinton [23] showed that a very deep and narrow DBN, with $\sim 3\cdot 2^n$ hidden layers of width $(n+1)$, can approximate any distribution on $\{0,1\}^n$ arbitrarily well. Le Roux and Bengio [11] improved this bound showing that $\sim \frac{2^n}{n}$ layers of width n suffice. Montúfar and Ay [15] improved that bound again to $\sim \frac{2^n}{2n}$. We are interested in the expressive power of DBNs which have less than $2^n - 1$ parameters and cannot approximate every probability distribution arbitrarily well. In [18] the maximal Kullback-Leibler approximation errors of RBMs were bounded from above by studying submodels of RBMs.

Definition 1. *A* submodel *of a DBN with layer widths $n_0,\dots,n_l$ is a set of probability distributions in $\Delta_{2^{n_0}-1}$ contained in* $\mathrm{DBN}(n_0,\dots,n_l)$.

Approaches to find explicit submodels of DBNs include studying

- The set $\mathrm{DBN}(n_0,\dots,n_l)$ as a mixture of conditional distributions with mixing distributions from the imbedded model $\mathrm{DBN}(n_1,\dots,n_l)$. This approach was proposed in [13] and used in [16] to study the expressive power of RBMs. In Section 2 we describe distributed mixtures of product distributions arising in layered networks.
- Models arising from *probability sharing* on RBMs. This idea has been used in [23, 11, 15] to study universal approximation of probability distributions by DBNs. To study submodels of DBNs, one imposes constraints on the number and type of sharing steps (the number and widths of the hidden layers). The submodels are sub-simplicial-complexes of Δ_{2^n-1}. In Section 3.2 we discuss certain faces of the probability simplex that can be represented by deep and narrow DBNs.
- The set of joint probability distributions on the states of all units of a DBN and their linear projections (by marginalization maps).
- Graphical submodels of the DBN such as RBMs and trees.

Understanding these items is helpful to lower bound the capabilities of deep belief networks.

The marginal probability distributions on the states of the visible units of a stochastic network with no direct connections between visible units, are mixtures of product distributions. We call a mixture *distributed* when the mixture components share parameters in some way. Distributed representations

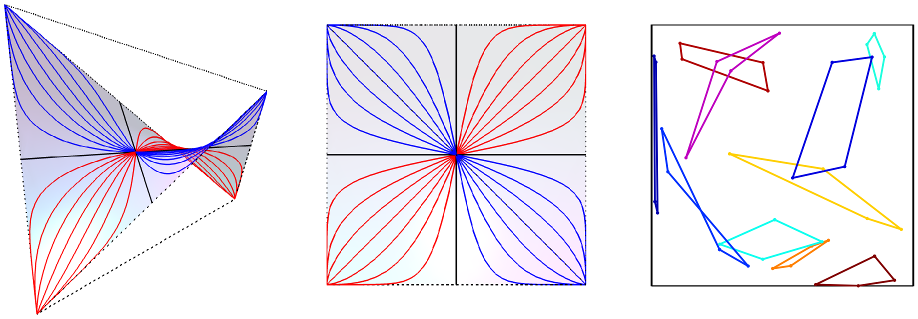

Figure 2: Left: Two-bit product distributions with straight lines αB, $\alpha \in \mathbb{R}$ as natural parameters for various choices of $B \in \mathbb{R}^2$. Middle: Linear projection of the left figure into the convex support of the two-bit independence model. Right: Linear projection of 10 $(2, 2)$-zonoset tuples of product distributions with random $W \in \mathbb{R}^{2\times 2}$, $B \in \mathbb{R}^2$.

have been discussed in [8, 1, 16]. Each layer of a DBN defines a distributed mixture of product distributions. Similarly, each layer of a *deep Boltzmann machine* (DBM) and a *directed RBM* define a distributed mixture of product distributions. A DBM is a layered network with undirected bipartitie connections between units in subsequent layers, see [20]. The DBM model is the set of marginal distributions on the states of the variables in the bottom layer. The model $\mathrm{RBM}^{\mathrm{dir}}_{n,m}$ is the set of visible distributions of a pair of layers of binary units with directed connections from the top layer to the bottom layer, including top and bottom bias weights, and without connections within each layer, as shown in Figure 1.

In Section 2 we discuss the mixtures of product distributions represented by layered networks. In Section 3 we study the geometry of the set of all stochastic transitions that can be realized by DBN layers. In Section 4 we derive upper bounds on the maximal and mean approximation errors of DBNs. Section 5 presents a discussion of our results. All formal proofs of mathematical statements are deferred to the Appendix.

2 Distributed mixtures of products and stochastic kernels

An exponential family is a set of probability distributions of the form $\mathcal{E}_V = \{p \propto \exp(f) \colon f \in V\}$, where V is an affine space of functions on the set of elementary events. The set of all strictly positive product distributions of n binary variables is an n-dimensional exponential family, denoted by $\mathcal{M}_n \subseteq \Delta_{2^n-1}$, with elements $p_B(v_1, \dots, v_n) = \prod_{i=1}^n p_{B_i}(v_i) = \exp(Bv)/Z_B$, $Z_B = \sum_{v\in\{0,1\}^n} \exp(Bv)$. Here $B \in \mathbb{R}^n$ is called the *natural parameter* vector of p_B. The convex support of this model is an n-dimensional hypercube with points in one-to-one correspondence with the points in the closure $\overline{\mathcal{M}_n}$ of $\mathcal{M}_n$. See [3].

The k-*mixture* of product distributions of n binary variables is

$$\mathcal{M}_{n,k} := \{\sum_{j=1}^{k} \lambda_j p^{(j)} \colon p^{(j)} \in \mathcal{M}_n, \lambda_j \geq 0 \;\forall j, \text{ and } \sum_{j=1}^{k} \lambda_j = 1\} \;. \tag{4}$$

This set has the dimension expected from counting parameters, $\dim(\mathcal{M}_{n,m}) = \min\{2^n - 1, mn + m - 1\}$, unless $n = 4$ and $m = 3$, see [4].

The marginal visible probability distributions of DBNs, DBMs, directed RBMs, and RBMs with n binary visible units and m binary units in the first hidden layer, all have the following form:

$$p(v) = \sum_{h\in\{0,1\}^m} p_{hW+B}(v)\, q(h) \quad \forall v \in \{0,1\}^n, \quad \text{where} \tag{5}$$

$$p_{hW+B}(v) = \frac{1}{Z_h} \exp((hW + B)v) \quad \forall v \in \{0,1\}^n, \quad \forall h \in \{0,1\}^m \;, \tag{6}$$

with $Z_h = \sum_{v\in\{0,1\}^n} \exp((hW + B)v)$, $W \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^n$, and q is a probability distribution on $h \in \{0,1\}^m$.

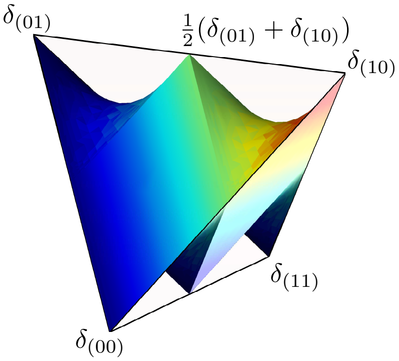

Figure 3: The set $u \cdot \mathcal{K}_{1,2} = \{u \cdot K_{W,B} \colon W \in \mathbb{R}^{1\times 2}, B \in \mathbb{R}^{1\times 2}\} \subset \Delta_3$, where $u = (1/2, 1/2)$.

The natural parameters $\mathcal{Z} = \{hW + B \colon h \in \{0,1\}^m\}$, with $W \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^n$, of the 2^m product distributions $\{p_{hW+B} \colon h \in \{0,1\}^m\}$, are a multiset (a set with repetitions allowed) of points in $\mathbb{R}^n$ called an (m,n)-*zonoset*. In the literature of polytopes the convex hull of a zonoset is known as *zonotope*.

Definition 2. *We call* $\{p_{hW+B} \colon h \in \{0,1\}^m\}$ *the* zonoset tuple *of product distributions associated to the zonoset* $\mathcal{Z} = \{hW + B \colon h \in \{0,1\}^m\}$*.*

The number of parameters of a zonoset tuple is $(m+1)n$, while $2^m n$ parameters are needed for describing an arbitrary tuple of 2^m product distributions. Any (m,n)-zonoset-tuple of product distributions is contained in an exponential subfamily of $\mathcal{M}_n$ of dimension $\min\{m,n\}$. Figure 2 illustrates zonoset tuples of product distributions on $\{0,1\}^2$.

We can view eq. (5) as a transition of the marginal distribution q on the states of the first hidden layer, to the visible distribution p, by a stochastic kernel:

$$p = q \cdot K_{W,B} \ , \tag{7}$$

where the kernel, called an (m,n)-*zonoset kernel*, is defined by the $2^m \times 2^n$-matrix with entries

$$K_{W,B}(h,v) := p_{hW+B}(v) \qquad \text{for all } h \in \{0,1\}^m \text{ and all } v \in \{0,1\}^n \ . \tag{8}$$

Thus a zonoset tuple is the rows of a zonoset kernel viewed as a set. Each $K_{W,B}$ is a (row) stochastic matrix describing a linear map

$$K_{W,B} \colon \Delta_{2^m-1} \to \operatorname{conv}\{K_{W,B}(h,\cdot)\}_h \subseteq \Delta_{2^n-1} \ ; \ q \mapsto p \cdot K_{W,B} \ .$$

We denote the set of all (m,n)-zonoset kernels by

$$\mathcal{K}_{m,n} := \{K_{W,B} \colon W \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^n\} \ .$$

We write $\overline{\mathcal{K}_{m,n}}$ for the set of all kernels that can be expressed as the limit of a sequence $K_{W_i,B_i} \in \mathcal{K}_{m,n}$, $i \in \mathbb{N}$.

The *input* distributions q in eq. (5) are restricted in different ways for each model:

- For DBNs $q \in \mathrm{DBN}(n_1,\dots,n_l)$, and $\mathrm{DBN}(n_0,\dots,n_l) = \mathrm{DBN}(n_1,\dots,n_l) \cdot \mathcal{K}_{n_1,n_0}$; in particular a DBN with layers of constant width is given by $\mathrm{RBM}_{n,n} \cdot \mathcal{K}_{n,n}^{l-2}$.
- For directed RBMs $q \in \mathcal{M}_m$, and $\mathrm{RBM}^{\mathrm{dir}}_{n,m} = \mathcal{M}_m \cdot \mathcal{K}_{m,n}$.
- For RBMs $q \in \{\frac{1}{Z}\sum_v \exp((hW+B)v + Ch) \colon C \in \mathbb{R}^m\}$.
- For DBMs $q \in \{\frac{Z_h}{Z}\sum_{h^2,\dots,h^l}\prod_{k=1}^{l-1}\exp((h^{k+1}W^{k+1} + B^k)h^k)\exp(B^l h^l)\}$.

In the case of RBMs and DBMs q is subject to "feedback" from the visible units and depends on W and B, while for DBNs and directed RBMs q is independent from these parameters. The 2^m product distributions p_{hW+B}, $h \in \{0,1\}^m$, which we summarized in the rows of $K_{W,B}$, however are the same for all these models. The smallest model which contains all models of the form $\mathcal{M} \cdot \mathcal{K}_{m,n}$

is the (m, n)-*zonoset mixture of products* (ZMP), defined by $\mathrm{ZMP}_{n,m} := \Delta_{2^m-1} \cdot \mathcal{K}_{m,n}$, or more explicitly:

$$\mathrm{ZMP}_{n,m} := \Big\{ \sum_{h\in\{0,1\}^m} \lambda_h \, p_{hW+B} \;\Big|\; \lambda_h \geq 0, \sum_h \lambda_h = 1, W \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^n \Big\} . \tag{9}$$

DBNs and DBMs are "cut out" from ZMPs by their specific constraints on the mixture weights $q(h)$. The mixture weights q of $\mathrm{DBN}(n_0, n_1, \dots, n_l)$ can be chosen arbitrarily and the model is equal to ZMP_{n_0,n_1} only if $\mathrm{DBN}(n_1, \dots, n_l)$ is a universal approximator on $\{0, 1\}^{n_1}$.

ZMPs are submodels of very large mixtures of products; $\mathrm{ZMP}_{n,m} \subseteq \mathcal{M}_{n,2^m}$, for all n and m. By results from [16], $\mathcal{M}_{n,2^m}$ is also the smallest mixture of products that contains $\mathrm{RBM}_{n,m}$ and thus $\mathrm{ZMP}_{n,m}$, when $4\lceil m/3\rceil \leq n$. On the other hand, each zonoset tuple shares the parameters W and B, and the largest mixture of products contained in a ZMP is possibly relatively small. The total number of parameters of $\mathrm{ZMP}_{n,m}$ is $(m + 1)n + 2^m - 1$. We note that $\mathcal{M}_{n,m+1} \subseteq \mathrm{ZMP}_{n,m}$ for all n and m, and $\mathcal{M}_{n,k} \nsubseteq \mathrm{ZMP}_{n,m}$ when $k > \frac{(m+1)n+2^m}{(n+1)}$ (by counting parameters).

Example 3. If the *input* q is a point measure δ_h, then the output is just the hth row $q \cdot K_{W,B} = p_{hW+B}$ of the kernel. In particular $\delta_h \cdot \mathcal{K}_{m,n} = \mathcal{M}_n$ for any h. If the input is the uniform distribution u on $\{0, 1\}^m$, then the output $p = q \cdot K_{W,B}$ is the arithmetic mean of a zonoset tuple. Figure 3 illustrates this set for one hidden and two visible units.

3 Geometry and combinatorics of zonoset kernels

A *face* or a *cylinder set* of the n-cube is a maximal set of binary vectors of length n with fixed values in a set of coordinates $I \subseteq [n]$. We write $[h^*_I] = \{h \in \{0, 1\}^n : h_I = h^*_I\}$ for the $(n - |I|)$-dimensional face with fixed values $h_i = h^*_i$ for all $i \in I$. We write $a \oplus_2 b$ for $a + b \mod 2$. Given a vector $h \in \{0, 1\}^m$ and a subset $I \subset [m]$, we write h_I for a vector in $\{0, 1\}^I$, or for the vector with entries $(h_I)_i = h_i$ if $i \in I$ and $(h_I)_i = 0$ if $i \notin I$. The support of a probability distribution p defined on a set $\mathcal{X}$ is $\mathrm{supp}(p) := \{x \in \mathcal{X} : p(x) > 0\}$.

We start showing that certain classes of kernels can be realized as zonoset kernels. Let $n = m$. Given any $p \in \Delta_{2^m-1}$, let $\mathrm{K}_p(h, v) := p(h \oplus_2 v)$. The rows of K_p are permuted versions of the probability distribution p. Figure 1 illustrates the set of all kernels K_p with p uniformly distributed on faces of $\{0, 1\}^3$. The *mixing times* of these kernels have been studied in the context of Markov chains on finite groups, see [21].

Proposition 4. *Let p be any product distribution with support on any face of $\{0, 1\}^n$ with fixed coordinates $I \subseteq [n]$, and let $\mathrm{K}_p(h, v) = p(h \oplus_2 v)$ for $v, h \in \{0, 1\}^n$. Then there is a zonoset kernel $K_{W,B} \in \overline{\mathcal{K}_{n,n}}$ with $K_{W,B}(h, v) = \mathrm{K}_p(h, h_{I^c} \oplus_2 v)$ for all $h, v \in \{0, 1\}^n$, and in particular:*

- $\mathrm{supp}(K_{W,B}(h, \cdot)) = \mathrm{supp}(\mathrm{K}_p(h, \cdot))$ *for all h,*
- $K_{W,B}(h, \cdot) = \mathrm{K}_p(h, \cdot)$ *for all h with* $\mathrm{supp}(h) \subseteq I$*, e.g., for $h = (0, \dots, 0)$,*
- *If p is uniformly distributed on a face of $\{0, 1\}^n$, then $K_{W,B} = \mathrm{K}_p$.*

The following propositions show that the set $\mathcal{K}_{m,n}$ has the dimension expected from parameter counting, and that its elements are generically full rank matrices.

Proposition 5. *The set of kernels $\mathcal{K}_{m,n}$ is a multigraded toric variety.*

Remark 6. Let V denote a sufficient statistics of the n-bit independence model, e.g., a matrix with columns the elements of $\{0, 1\}^n$, and let H be the $(m + 1) \times 2^m$-matrix with columns $\{(1, h)\}_{h\in\{0,1\}^m}$. Consider the exponential family $\mathcal{E}_{H\otimes V}$ with sufficient statistics $H \otimes V$ on $\mathcal{X} = \{0, 1\}^{n+m}$. Let $\mathcal{X}_h = \{(v, h') \in \mathcal{X} : h' = h\} \cong \{0, 1\}^n$ for all h. For each $p \in \mathcal{E}_{H\otimes V}$ there is a $K_{W,B} \in \mathcal{K}_{m,n}$ (and vice versa) with $p(\cdot|\mathcal{X}_h) = K_{W,B}(h, \cdot)$ $\forall h \in \{0, 1\}^m$. In particular, $\dim(\mathcal{K}_{m,n}) = (m + 1)n$, as expected from counting parameters.

Proposition 7. *Assume that all rows of $W \in \mathbb{R}^{m\times n}$ are multiples of the same vector $C \in \mathbb{R}^n$, i.e., $W = (\alpha_k C)^m_{k=1}$. For almost every C and $(\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$ the kernel $K_{W,B}$ is totally non-vanishing, i.e., all its minors are non-vanishing.*

Proposition 8. *For any n and m the kernels $K_{W,B} \in \mathcal{K}_{m,n}$ are full rank for almost all choices of W and B. In particular, almost every zonoset kernel $K_{W,B} \in \mathcal{K}_{m,n}$ is injective when $m \leq n$, and* $\dim(\Delta_{2^m-1} \cdot K_{W,B}) = \min\{2^m - 1, 2^n - 1\}$.

Example 9. Consider an RBM and a directed RBM, both with m hidden and n visible binary units, $m \leq n$. For almost all fixed choices of W and B, the sets of probability distributions $\{\sum_h p_{hW+B}(v)\, \frac{Z_h}{Z} p_C(h) \colon C \in \mathbb{R}^m\}$ and $\{\sum_h p_{hW+B}(v)\, p_C(h) \colon C \in \mathbb{R}^m\}$ represented respectively by the two models as the bias of the hidden units vary, are almost everywhere different from each other (their intersection has dimension strictly less than m). When training DBNs, the DBN modules (directed RBMs) are commonly treated as RBMs. By this example, the probability distributions that can possibly be represented by the DBN modules almost never match the trained RBM distributions.

The binary vectors $\{0,1\}^n$ are the vertices of the n-dimensional unit hypercube. We call *edge* a pair $\{x, y\} \subset \{0,1\}^n$ with $d_H(x,y) = 1$, where $d_H(x,y) := |\{i \in [n] \colon x_i \neq y_i\}|$ denotes the Hamming distance between x and y.

Proposition 10. *Each of the following tuples of product distributions can be realized as a subset of rows of a zonoset kernel with an appropriate choice of W and B:*

1. *If $\mathcal{C} \subset \{0,1\}^m$, $|\mathcal{C}| = m+1$ are affinely independent vectors (over $\mathbb{R}^m$), e.g., $\mathcal{C}$ is a Hamming ball of radius 1 in $\{0,1\}^m$, then $\{p_{hW+B}\}_{h \in \mathcal{C}}$ are any $m+1$ product distributions.*

2. *Let $\mathcal{C}$ be a K-dimensional face of the m-cube, $K \leq m$. The set $\{p_{hW+B}\}_{h\in\mathcal{C}}$ contains the uniform distributions on the (nonempty) intersections of any K faces of the n-cube.*

3. *Let $\lambda \subseteq [n] := \{1, \dots, n\}$ and $\Lambda \subseteq [m]$ with $|\lambda| = |\Lambda| = K$. Let $\mathcal{C}$ be a K-face of the m-cube with free coordinates Λ. p_{hW+B} is the uniform distribution on $\{x \colon x_\lambda = h_\Lambda\}$ for all $h \in \mathcal{C}$. Note that $\{x \colon x_\lambda = h_\Lambda\}_{h\in\mathcal{C}}$ is a partition of $\{0,1\}^n$ into blocks of cardinality 2^{n-K}.*

4. *Let $m = n$ and let $\{h^{i+}, h^{i-}\}$, $i = 1, \dots, m$ be m disjoint edges of the m-cube. $p_{h^{i+}W+B}$ is any distribution supported on the edge $\{h^{i+}, h^{i+} \oplus_2 + \mathbf{e}_i\}$, and $p_{h^{i-}W+B}$ is any distribution supported on the edge $\{h^{i-}, h^{i-} \oplus \mathbf{e}_i\}$, for all $i \in [m]$. Moreover, $p_{hW+B} = \delta_h$ for all $h \notin \cup_i\{h^{i+}, h^{i-}\}$. (This statement in fact summarizes [11, Theorems 1 and 2]).*

Corollary 11. *The model* $\mathrm{DBN}(n,m,m)$ *contains the mixture model $\mathcal{M}_{n,m+1}$. In contrast,* $\mathrm{RBM}_{n,m}$ *does not contain $\mathcal{M}_{n,m+1}$, in general.*

3.1 Patterns of modes in zonoset tuples

In the following we elaborate on the sets of modes that can be realized jointly by rows of zonoset kernels, slightly extending results on RBMs and mixtures of products shown in [16].

A *mode* of a probability distribution $p \in \Delta_{2^n-1}$ is point $x \in \{0,1\}^n$ such that $p(x) > p(y)$ for all y with $d_H(x,y) = 1$. The set of *strong modes* [16] of p is $\{x \in \{0,1\}^n \colon p(x) \geq \sum_{d_H(x,y)=1} p(y)\}$. We denote by $\mathcal{H}_\mathcal{C} \subseteq \Delta_{2^n-1}$ the set of probability distributions with strong modes $\mathcal{C}$. An *n-bit code* $\mathcal{C}$ is just a subset of $\{0,1\}^n$. The *minimum distance* of $\mathcal{C}$ is defined as $\min\{d_H(x,y) \colon x, y \in \mathcal{C}, x \neq y\}$. Given a sign vector $s \in \{-,+\}^n$, the *s-orthant* of $\mathbb{R}^n$ is the set of all vectors in $\mathbb{R}^n$ with sign s. We identify sign vectors $\{-,+\}^n$ and binary vectors $\{0,1\}^n$ via $- \mapsto 0$ and $+ \mapsto 1$.

Proposition 12.

1. *Let $\mathcal{C} \subset \{0,1\}^n$ be a code of minimum distance two. If the model* $\mathrm{ZMP}_{n,m}$ *contains a probability distribution with strong modes $\mathcal{C}$, then there is an (m,n)-zonoset with a point in every s-orthant of $\mathbb{R}^n$, $s \in \mathcal{C}$.*

2. *If* $\mathrm{ZMP}(n_0, n_1)$ *contains probability distributions with* 2^{n_0-1} *strong modes, then* $n_1 \geq n_0 - 1$. *In fact $n_1 \geq n_0$, when n_0 is odd and larger than one.*

3. *If* $\mathrm{ZMP}(n_0, n_1)$ *is a universal approximator of distributions from $\Delta_{2^{n_0}-1}$ with $n_0 \geq 7$, then $n_1 \geq n_0$.*

In particular, when $n_0 \geq 3$, the DBNs with layers of widths $n_0 > n_1 > \cdots > n_l$ cannot represent distributions with 2^{n_0-1} strong modes. If $\mathrm{DBN}(n_0, n_1, \dots, n_l)$ is a universal approximator with $n_1 = n_0 - 1$ (and $n_0 \leq 6$), then $\mathrm{DBN}(n_1, n_2, \dots, n_l)$ is also a universal approximator.

A *linear threshold code* (LTC) is a subset of $\{0,1\}^n$ that corresponds to the sign vectors of the points of a zonoset in $\mathbb{R}^n$. Equivalently, an LTC is an admissible multi-labeling of the vertices of a hypercube by a collection of linear threshold functions.

Proposition 13. *Let $\mathcal{C} \subseteq \{0,1\}^n$, $|\mathcal{C}| = 2^m$ be a code of minimum distance two. Then both $u \cdot \mathcal{K}_{n,m}$ and $\Delta_{2^m-1} \cdot \mathcal{K}_{n,m}$ contain a distribution with strong modes $\mathcal{C}$ iff $\mathcal{C}$ is a linear threshold code.*

Proposition 14.

- *If $4\lceil m/3 \rceil \leq n$, then $u \cdot \mathcal{K}_{n,m} \cap \mathcal{H}_{n,2^m} \neq \emptyset$ and $\mathcal{M}_{n,k} \supseteq u \cdot \mathcal{K}_{n,m}$ iff $k \geq 2^m$.*
- *If $4\lceil m/3 \rceil > n$, then $u \cdot \mathcal{K}_{n,m} \cap \mathcal{H}_{n,L} \neq \emptyset$, where $L := \min\{2^l + m - l, 2^{n-1}\}$, $l := \max\{l \in \mathbb{N} \colon 4\lceil l/3 \rceil \leq n\}$, and $\mathcal{M}_{n,k} \supseteq u \cdot \mathcal{K}_{n,m}$ only if $k \geq L$.*

3.2 Submodels of DBNs from probability sharing

The idea of this subsection is to propagate the probability mass of distributions generated by the top RBM of a DBN across the network, in order to learn something about the visible probability distributions at the bottom. This can be accomplished by describing the products of kernels $\mathcal{K}_{n_{l-1},n_{l-2}} \cdot \mathcal{K}_{n_{l-2},n_{l-3}} \cdots \mathcal{K}_{n_1,n_0}$. For simplicity we shall consider layers of same width as the visible layer, n. In this case the propagation can be interpreted as a process in the graph of a hypercube.

A kernel realizes sharing of probability from a state $a \in \{0,1\}^n$ to a state $b \in \{0,1\}^n$ if its ath row has non-vanishing bth entry. It is possible to share probability from a to a collection of states $b^{(1)}, \dots, b^{(s)}$ in arbitrary ratios by a product of l kernels iff the ath row of a product of kernels in $\mathcal{K}^l_{n,n}$ can be made an arbitrary distribution on $b^{(1)}, \dots, b^{(s)}$. In particular, since all rows of zonoset kernels are product distributions, probability sharing from one state to more than two states, in arbitrary rations, is not possible in one single DBN layer.

An *l-path* on the graph of the n-cube is a list S of l vectors in $\{0,1\}^n$ with subsequent elements differing in at most one bit, $S_1, \dots, S_l \in \{0,1\}^n$, $d_H(S_k, S_{k+1}) \leq 1$. An n-bit *Gray code* of length l is a special l-path with different subsequent elements. The transition sequence T of a path is the list of bit-indices where the subsequent elements differ from each other (possible empty).

Let $\mathcal{S}(\mathrm{RBM}_{n,m})$ denote the collection of support sets of all faces of the probability simplex Δ_{2^n-1}, which are contained in $\overline{\mathrm{RBM}_{n,m}}$. It is known that any union of $(m+1)$ edges of the n-cube is is in $\mathcal{S}(\mathrm{RBM}_{n,m})$, see [15, Theorem 1]. Consider some $R \in \mathcal{S}(\mathrm{RBM}_{n,n})$ and a collection of l-paths S^i starting from R, such that at any time $1 \leq t \leq l-1$ two paths change the same bit only if they are visiting neighboring points. We denote the collection of all such sets by

$$\mathbb{S}^l_n := \left\{ \cup_{i \in R} S^i \middle| \cup_i S^i_1 = R \in \mathcal{S}(\mathrm{RBM}_{n,n}), T^i_t \neq T^j_t \text{ unless } d_H(S^i_t, S^j_t) = 1 \right\} . \qquad (10)$$

The following result generalizes [11, Lemma 1, Theorems 1 and 2] to DBNs with any number of layers of constant width:

Lemma 15. *The model $\mathrm{DBN}(n, \dots, n)$ with l hidden layers contains any probability distribution with support in an element of $\mathbb{S}^l_n$.*

For some elements of $\mathbb{S}^l_n$ we find an explicit description:

Proposition 16. *If $n \geq N(2^k + k + 1)$ and $l \geq 2^{2^k}$ for some $k \in \mathbb{N}$, then $\mathbb{S}^l_n$ contains the union of N arbitrary $(2^k + k + 1)$-dimensional faces of the n-cube with disjoint free coordinates. In particular, when $l \geq 2^n/2(n - \log(n))$, the entire state space $\{0,1\}^n$ is an element of $\mathbb{S}^l_n$.*

4 Expressive power and approximation errors of DBNs

In this section we describe some submodels of DBNs explicitly, and use them to bound the approximation errors of DBNs from above.

Let $\varrho = \{A_1, \dots, A_K\}$ be a partition of $\{0,1\}^n$. The *partition model* $\mathcal{M}_\varrho$ is the set of all probability distributions with $p(x) = p(y)$ whenever x and y belong to the same block A_i of the partition ϱ.

The following collects some results shown in the previous section:

Theorem 17. *Let $l \in \mathbb{N}$. Let k be the largest natural number for which $l - 1 \geq 2^{2^k}$, and let $K = 2^k + k + 1 \leq n$. The model* $\mathrm{DBN}(n, \dots, n)$ *with l hidden layers contains:*

- *Any $p \in \Delta_{2^n-1}$ with support contained in an element of $\mathbb{S}_n^l$.*
- *Any partition model $\mathcal{M}_\varrho$ with partition $\varrho = \{[y_\lambda]\}_{y_\lambda \in \{0,1\}^K}$, $\lambda \subseteq [n], |\lambda| = K$.*

If $K \geq n$, then the DBN is a universal approximator, which is consistent with [15, Theorem 1].

The Kullback-Leibler divergence from a point p to a model $\mathcal{M}$ in Δ_{2^n-1} is defined as $D(p\|\mathcal{M}) := \inf_{q \in \mathcal{M}} D(p\|q)$, where $D(p\|q) := \sum_{x \in \{0,1\}^n} p(x) \log \frac{p(x)}{q(x)}$ is the divergence from p to q. The *maximal KL-divergence* [12, 18] from a partition model $\mathcal{M}_\varrho$ with 2^K blocks of cardinalities 2^{n-K}, as given in the second item of Theorem 17, is $\max_{p \in \Delta_{2^n-1}} D(p\|\mathcal{M}_\varrho) = (n - K)$, see [18, Corollary 3.1]. The *Dirichlet prior* on Δ_{2^n-1} with concentration parameter $\boldsymbol{\alpha} = (\alpha_x)_{x \in \{0,1\}^n}$ is $\mathrm{Dir}_{\boldsymbol{\alpha}}(p) := \frac{1}{\sqrt{2^n}} \frac{\Gamma(\sum_x \alpha_x)}{\prod_x \alpha_x} \prod_x p(x)^{\alpha_x - 1}$ for all $p \in \Delta_{2^n-1}$, whereby the sums and products are over $x \in \{0,1\}^n$. If p is drawn from this prior, then the expected approximation error is, see [17, Theorem 4]:

$$\mathbb{E}[D(p\|\mathcal{M}_\varrho)] = (n-K)\ln(2) + \sum_{x \in \{0,1\}^n} \frac{\alpha_x}{\sum_y \alpha_y} h(\alpha_x) - \sum_{j=1}^{2^K} \frac{\sum_{x \in A_j} \alpha_x}{\sum_y \alpha_y} h(\sum_{x \in A_j} \alpha_x) \,, \quad (11)$$

where $h(k) := 1 + \frac{1}{2} + \cdots + \frac{1}{k}$ denotes the kth *harmonic number*.

The approximation error of a DBN is bounded from above by the approximation error of any of its submodels. If we use any of the partition models with 2^K blocks of cardinalities 2^{n-K}, we get:

Theorem 18. *Consider a DBN with l hidden layers of width n.*

- *The maximal KL-approximation error of this model is bounded from above by*

$$\max_{p \in \Delta_{2^n-1}} D(p\|\,\mathrm{DBN}) \leq n - K \,, \quad \textit{where } K = 2^k + k + 1 = \log(2l\log(l)).$$

- *The expected KL-approximation error is bounded from above by eq.* (11)*. In particular, if p is drawn uniformly at random from the probability simplex Δ_{2^n-1}, then the expected divergence* $\mathbb{E}[D(p\|\,\mathrm{DBN})]$ *is bounded from above by* $1 + \ln(2^{n-K}) - h(2^{n-K})$.

5 Discussion

Deep belief networks generate mixtures of tuples of product distributions whose parameters are projections of hypercubes' vertices (zonosets), described by very few shared parameters. We cast these tuples of product distributions as the rows of stochastic matrices (zonoset kernels), and studied properties such as their rank, symmetries, and combinatorics.

This analysis exposes similarities of DBNs and DBMs, and shows possible ways of defining distributed mixtures of products; e.g., as $\mathcal{E} \cdot \mathcal{K}$, with a low-dimensional model $\mathcal{E} \in \Delta_{2^m-1}$, and a family of kernels $\mathcal{K}$. The rows of each kernel in the family $\mathcal{K}$ can be chosen as product distributions with parameters equal to the projected vertices of a hypercube, or the projected vertices of any other low-dimensional polytope. In contrast, standard, unrestricted mixtures of products, correspond to projected vertices of (high-dimensional) simplices.

Kernels are helpful for understanding probability sharing in layered networks. We showed explicit classes of probability distributions than can be learned by DBNs depending on the number of hidden layers that they contain. Various submodels of RBMs with k parameters, such as unions of partition models, can be learned by deep and narrow DBNs with k parameters. We showed that the maximal approximation error of narrow DBNs is not larger than the upper bounds on the approximation errors of RBMs with the same number of parameters shown in [18].

Furthermore, we bounded the expected approximation error of DBNs from above. Our bounds are with respect to Dirichlet priors. These priors do not only have technical advantages, but are a canonical choice when no information is availble about the real distribution of the targets. It could be interesting to consider other priors in future work. We note in particular, that the exact expected error formula from Theorem 18 item 2, eq. 11, can be integrated over an hyperprior of interest.

The approximation error bounds from Theorem 18 can possibly be improved by taking into account the totality of DBN submodels described in this paper, instead of just partition models. It is worth mentioning that any DBN which is a graphical supermodel of $\mathrm{DBN}(n_0, n_0 - 1, n_0 - 2, \ldots, 1)$ has the general Markov model corresponding to any tree on n_0 leaves as a graphical submodel. That is, this DBN contains the union of all such tree models. Furthermore, DBNs often contain Hadamard products of trees as well, so it is possible to study their dimension by *tropicalization* [19].

Acknowledgments

This work is supported in part by DARPA grant FA8650-11-1-7145.

References

[1] Y. Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, 2009.

[2] Y. Bengio and O. Delalleau. On the expressive power of deep architectures. In J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, editors, *ALT*, volume 6925 of *Lecture Notes in Computer Science*, pages 18–36. Springer, 2011.

[3] L. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

[4] M. V. Catalisano, A. V. Geramita, and A. Gimigliano. Secant varieties of $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ (n-times) are not defective for $n \geq 5$. *J. Algebraic Geometry*, 20:295–327, 2011.

[5] M. A. Cueto, J. Morton, and B. Sturmfels. Geometry of the restricted Boltzmann machine. In M. A. G. Viana and H. P. Wynn, editors, *Algebraic methods in statistics and probability II, AMS Special Session*, volume 2. American Mathematical Society, 2010.

[6] Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using 2-layer networks. In *Advances in Neural Information Processing Systems 4*, pages 912–919. 1992.

[7] R. Hartshorne. *Algebraic Geometry*. Graduate Texts in Mathematics. Springer, 1977.

[8] G. E. Hinton. Products of experts. In *Proceedings 9-th ICANN*, volume 1, pages 1–6, 1999.

[9] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[10] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[11] N. Le Roux and Y. Bengio. Deep belief networks are compact universal approximators. *Neural Computation*, 22:2192–2207, 2010.

[12] F. Matúš and N. Ay. On maximization of the information divergence from an exponential family. In *Proceedings of the WUPES'03*, pages 199–204. University of Economics, Prague, 2003.

[13] G. Montúfar. Mixture models and representational power of RBMs, DBNs, and DBMs. *Deep Learning and Unsupervised Feature Learning Workshop—NIPS'10*, 2010.

[14] G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *To appear in Kybernetika*, 2012. Preprint available at `http://arxiv.org/abs/1008.0204`.

[15] G. Montúfar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 23(5):1306–1319, 2011.

[16] G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *Deep Learning and Unsupervised Feature Learning Workshop—NIPS'12*, 2012. Preprint available at `http://arxiv.org/abs/1206.0387`.

[17] G. Montúfar and J. Rauh. Scaling of model approximation errors and expected entropy distances. In *To appear in WUPES'12*. University of Economics, Prague, 2012. Preprint available at `http://arxiv.org/abs/1207.3399`.

[18] G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 415–423. 2011.

[19] L. Pachter and B. Sturmfels. Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16132–16137, Nov. 2004.

[20] R. Salakhutdinov and G. Hinton. Deep Boltzmann machines. *Proceedings of the international conference on artificial intelligence and statistics*, 5(2):448–455, 2009.

[21] L. Saloff-Coste. *Probability on Discrete Structures*, chapter Random Walks on Finite Groups. Encyclopaedia of Mathematical Sciences. Springer Verlag, 2003.

[22] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Symposium on Parallel and Distributed Processing*, 1986.

[23] I. Sutskever and G. E. Hinton. Deep narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20:2629–2636, 2008.

Proofs

Geometry and combinatorics of zonoset kernels

Proof of Proposition 4. The kernel $K_{h_I^*} \equiv K_{[u_{h_I^*}]}$ has rows equal to the indicator functions of $h \oplus_2 [h_I^*]$, $h \in \{0,1\}^n$, multiplied by the constant $2^{-(n-|I|)}$. Note that $[h_I^*] = \mathbf{e}_i \oplus_2 [h_I^*]$ for all $i \in [n] \setminus I$. For each $v_{[n]\setminus I} \in \{0,1\}^{[n]\setminus I}$, the sets $(v_I, v_{[n]\setminus I}) \oplus_2 [h_I^*]$, $v_I \in \{0,1\}^I$ partition $\{0,1\}^n$ into $2^{|I|}$ cylinder sets. The connection weights $W(i,j) = \alpha(-h_i^* + \frac{1}{2})\delta_i(j)\mathbb{1}_I(j)$ and the bias weights $B(j) = -\alpha\frac{1}{2}(-h_j^* + \frac{1}{2})\mathbb{1}_I(j)$ produce the kernel

$$K_{W,B}(h,v) = \exp(\alpha\frac{1}{2}(-h^*_{I\cap \operatorname{supp} h} + \frac{1}{2}\mathbb{1}_{I\cap\operatorname{supp} h} + h^*_{I\setminus \operatorname{supp} h} - \frac{1}{2}\mathbb{1}_{I\setminus \operatorname{supp} h})v)/Z.$$

The limit $\lim_{\alpha\to\infty} K_{W,B}$ is equal to $K_{[h_I^*]}$. To complete the proof we add the natural parameter vector C_{I^c} of p to the previously defined bias vector B. Then $K_{hW+B+C_{I^c}}$ satisfies the claims. □

Proof of Proposition 5. Replacing the parameters W_{ij}, B_j with their exponentials ω_{ij} and β_j, we obtain a multigraded monomial map $Q : \mathbb{C}^{nm+n} \to \prod_{i=1}^m \mathbb{P}^{2^n-1}$; $q_{h,v} = \prod_{j=1}^n \beta_j^{v_j} \prod_{i=1}^m \omega_{ij}^{h_i v_j}$. The Zariski closure of the image of this map is a multigraded toric variety inside a product of 2^m, $(2^n - 1)$-dimensional projective spaces, one for each hidden state. This variety is cut out by a multigraded monomial ideal generated by the multigraded binomials appearing in the kernel. □

Proof of Proposition 7. The rows of $K_{W,B}$ are the product distributions with natural parameters the zonoset generated by W and B. For assessing the rank of $K_{W,B}$ we may neglect the normalizing constants, and consider the matrix $\tilde{K}_{W,B}$ with rows $(\exp((hW + B)v))_{v\in\{0,1\}^n}$, $h \in \{0,1\}^m$. Furthermore, for any B with finite entries, the rank of $\tilde{K}_{W,B}$ and $\operatorname{diag}(\exp(-Bv))_v \cdot \tilde{K}_{W,B} = \tilde{K}_{W,\mathbf{0}}$ is equal.

Given the assumptions, the zonoset $\mathcal{Z} = \{hW + B \colon h \in \{0,1\}^m\}$ is contained in a straight line $\mathcal{Z} = \{\lambda_j C + B\}_{j=1}^{2^m}$, whereby the numbers $\lambda_j \in \mathbb{R}$ are all different from each other, for almost all $(\alpha_k)_k \in \mathbb{R}^m$. Let $(t_1, \dots, t_{2^n}) := (\exp(Cv))_{v\in\{0,1\}^n}$. Note that $t_i > 0$ for all i, and all t_i are different from each other, for almost all $C \in \mathbb{R}^n$. The rank of $\tilde{K}_{W,B}$ is equal to the rank of $(t_i^{\lambda_j})_{i,j}$, which, after some permutation of rows and columns, is a generalized Vandermonde matrix, known to be totally positive. Hence $\det(K_{W,B}(h,v))_{h\in H, v\in V} \neq 0$ for all $H \subseteq \{0,1\}^m$ and $V \subseteq \{0,1\}^n$, as claimed. □

Proof of Proposition 8. 1) First note that there is an open subset $\Omega \subset \mathbb{R}^{m\times n} \times \mathbb{R}^n$ of parameters W, B for which the kernels $K_{W,B}$ are full rank: Assume that the zonoset $\{hW + B \colon h \in \{0,1\}^m\}$ intersects 2^m orthants of $\mathbb{R}^n$, e.g., $W = I_n$ and $B = \frac{1}{2}(1, \dots, 1)$. Then $K_{\alpha W, \alpha B}$ is full rank for all α larger than some $a \in \mathbb{R}$, because for $\alpha \to \infty$ each row of $K_{\alpha W,\alpha B}$ converges to a different point measure. 2) Now, by Proposition 5 $\mathcal{K}_{m,n}$ is a (toric, irreducible) variety for all $m, n \in \mathbb{N}_0$. Let $l = \min\{m,n\}$. The set H of rank-deficient matrices in $\mathbb{C}^{2^l\times 2^l}$, or in $\prod_{i=1}^m \mathbb{P}^{2^n-1}$, is a hypersurface cut out by the vanishing of the determinant (which is a homogeneous polynomial on the matrix entries). Since $\mathcal{K}_{l,l} \not\subseteq H$, by [7, Proposition 7.1], every irreducible component of $\mathcal{K}_{l,l} \cap H$ has dimension $\dim(\mathcal{K}_{l,l}) - 1$. This is also an upper bound for the dimension of the real part of the set of rank-deficient kernels in $\mathcal{K}_{l,l}$. □

Proof of Example 9. Both models have the same zonoset kernels. For any choice of W and B, the set of inputs of the directed RBM are the product distributions $\mathcal{M}_m$. The set of inputs of the RBM are the distributions $q(h) = \frac{Z_h}{Z} \cdot \exp(Ch)$, which are product distributions iff $z(h) = \frac{Z_h}{\sum_h Z_h} \in \Delta_{2^m-1}$ is a product distribution. In [5] it is shown that $\{\frac{1}{Z}\sum_v \exp(hWv+Bv)\colon W, B\}$ is a set of dimension $mn + n$ when $n...m$. Since $K_{W,B}$ is injective, each output has a unique preimage. □

Proof of Corollary 11. By item 1 of Proposition 10, the set of distributions $q \in \Delta_{2^m-1}$ with support on a radius-one Hamming ball is mapped by $\mathcal{K}_{m,n}$ into the $(m+1)$-mixture of product distributions $\mathcal{M}_{n,m+1}$. The claim follows using that $\mathrm{RBM}_{n,m}$ contains any p with $|\operatorname{supp}(p)| \leq m+1$, see [15]. That RBMs do not contain the mixture model is a result from [16]. □

Patterns of modes in zonoset tuples

Proof of Proposition 12.

1. The first item follows from [16, Theorems 3 and 11].

2. For the first part: The number of strong modes of a mixture of k binary product distributions is at most k [16, Theorem 3]. For the second part: If n is odd and larger than one, then the smallest mixture of binary product distributions whose natural parameters are a zonoset and which approximates $u_{Z_{\pm,n}}$ arbitrarily well, has a zonoset generated by at least n vectors. See [16, Proposition 14].

3. The first part of the third item follows from parameter counting: The model $\sum_h \frac{1}{Z}\exp((hW + B)v)p(h)$, $p \in \Delta_{2^{n-1}-1}$ has a total of $n^2 + n + 2^{n-1} - 1$ parameters. This number is smaller than $\dim(\Delta_{2^n-1}) = 2^n - 1$ when $n \geq 7$. For the second part: Any mixture of binary product distributions which approximates some p with support $Z_{\pm,n}$ arbitrarily well, mixes the 2^{n-1} Dirac distributions δ_v, $v \in Z_{\pm,n}$, see [14]. Hence if $\mathrm{DBN}(n_0, \ldots, n_l)$ approximates any distribution p with support $Z_{\pm,n}$ arbitrarily well, then the mixture weights (distributions from $\mathrm{DBN}(n_1^l)$) approximate $p|_{Z_{\pm,n}}$ arbitrarily well. □

Proof of Proposition 13. This is a direct consequence of the analysis from [16]. □

Proof of Proposition 14. The proof of the first item follows the lines of the proof of [16, Theorem 32]. For the second item, note that if $\mathrm{DBN}(n, m, \ldots)$ can represent some p, then $\mathrm{DBN}(n, m + 1, \ldots)$ can represent $\lambda p + (1 - \lambda)\delta_x$ for any $x \in \{0, 1\}^n$ for some $0 < \lambda < 1$. □

Proof of Theorem 15. This result is a straightforward generalization of [11, Lemma 1, Theorems 1 and 2]. The elements of $\mathbb{S}_n^l$ meet the conditions of these lemma and theorems by definition. □

Submodels of DBNs from probability sharing

Proof of Proposition 16.

1. This follows immediately from [15, Lemma 4]. Any sub-DBN with layers of width $(n - R)$ is contained in the DBN with layers of width n. The distribution on the states of the remaining R visible nodes can be set to a point measure.

2. This follows from a similar argument as the first item. Any set of cardinality $(n + 1)$ is an S-set of $\mathrm{RBM}_{n,n}$. □

Proof of Proposition 10.

1 If $\mathcal{C} = \{h^{(0)}, \ldots, h^{(m)}\}$ are affinely independent, then $\{h^{(1)} - h^{(0)}, \ldots, h^{(m)} - h^{(0)}\}$ are linearly independent and can be mapped by W to an arbitrary set $\{W'_1, \ldots, W'_m\} \subset \mathbb{R}^n$. Choosing $B = B' - h^{(0)}W$, we can make $\{hW + B\colon h \in \mathcal{C}\}$ be arbitrary vectors $B', W'_1, \ldots, W'_m$, and so, $\{p_{hW+B}\colon h \in \mathcal{C}\}$ is an arbitrary set of $m + 1$ product distributions.

2 Any h can be identified with its support set. $p_{\{i\}}$ $i \in [m]$ are m uniform distributions on arbitrary faces F_i of the n-cube. p_λ is uniformly distributed with support $\operatorname{argmax}(\sum_{i\in\lambda} e_{F_i})$. E.g., if $F_\lambda := \cap_{i\in\lambda} F_i \neq \emptyset$, then $\operatorname{supp}(p_\lambda) = F_\lambda$.

3 This follows from the choice $W_{:,\lambda} = \alpha I_\lambda$, the identity matrix, and $W_{:,[n]\setminus\lambda} = 0$.

4 Consider any $l \in [n]$. Consider a pair of vectors $\{x, y\}$ which is an edge of $\{0,1\}^m$. Let $r \in [m]$ be the entry where they differ. Let $s \in [m]$ be arbitrary. Denote by $\hat{x}$ the vector $\hat{x}_i = x_i \, \forall i \neq r, s$ and $\hat{x}_r = 0$, $\hat{x}_s = 0$. Denote by $\mathbf{e}_i$ the vector with one 1 at the position i and zeros else. By $\mathbb{1}$ the vector of ones. Choosing

$$\begin{aligned} W_{:,l} &= \omega(2\hat{x} - \hat{\mathbb{1}} + (1 - 2x_s)m\mathbf{e}_s + (p - q)\mathbf{e}_r) \\ b_l &= -\omega(|\operatorname{supp}(x)| - 1 + x_s m) + q \end{aligned}$$

yields in the limit $\omega \to \infty$ that $P(v_l = h_s | h \neq x, y) = 1$, $P(v_l = 1 | h = x) = p$, and $P(v_l = 1 | h = y) = q$, i.e.,

$$\begin{aligned} P(v_l | h \neq x, y) &= \delta_{h_s}(v_l) \\ P(v_l | h = x) &= p(v_l) \\ P(v_l | h = y) &= q(v_l) \ . \end{aligned}$$

Consider the case $m = n$. Let $\{x^i, y^i\}_{i=1}^m$ be m disjoint edges of $\{0,1\}^m$. Let $s^i = i \; \forall i \in [m]$. Consider any $l \in [n]$. From the above discussion we get

$$P(v | h = x^l) = \prod_{i=1}^{n} P(v_i | x^l) = \prod_{i \neq l} \delta_{x^l_{s^i}}(v_i) \cdot p^l(v_l) \ , \tag{12}$$

which is an arbitrary distribution with support on the edge given by fixing $v_i = x_i^l \; \forall i \neq l$. For $h \notin \cup_{i=1}^m \{x^i, y^i\}$ and $s^i = i \, \forall i$ we get

$$P(v | h \neq x^l, y^l \, \forall l) = \prod_{i=1}^{n} P(v_i | h) = \prod_{i} \delta_{h_{s^i}}(v_i) = \delta_h(v) \ , \tag{13}$$

which is the point measure on $\{v = h\}$. □

Example 19. Figure 4 gives an example of zonoset kernels $K_{W,B} = \mathrm{K}_p$ in $\mathcal{K}_{4,4}$ for p the uniform distributions on faces of $\{0,1\}^4$.

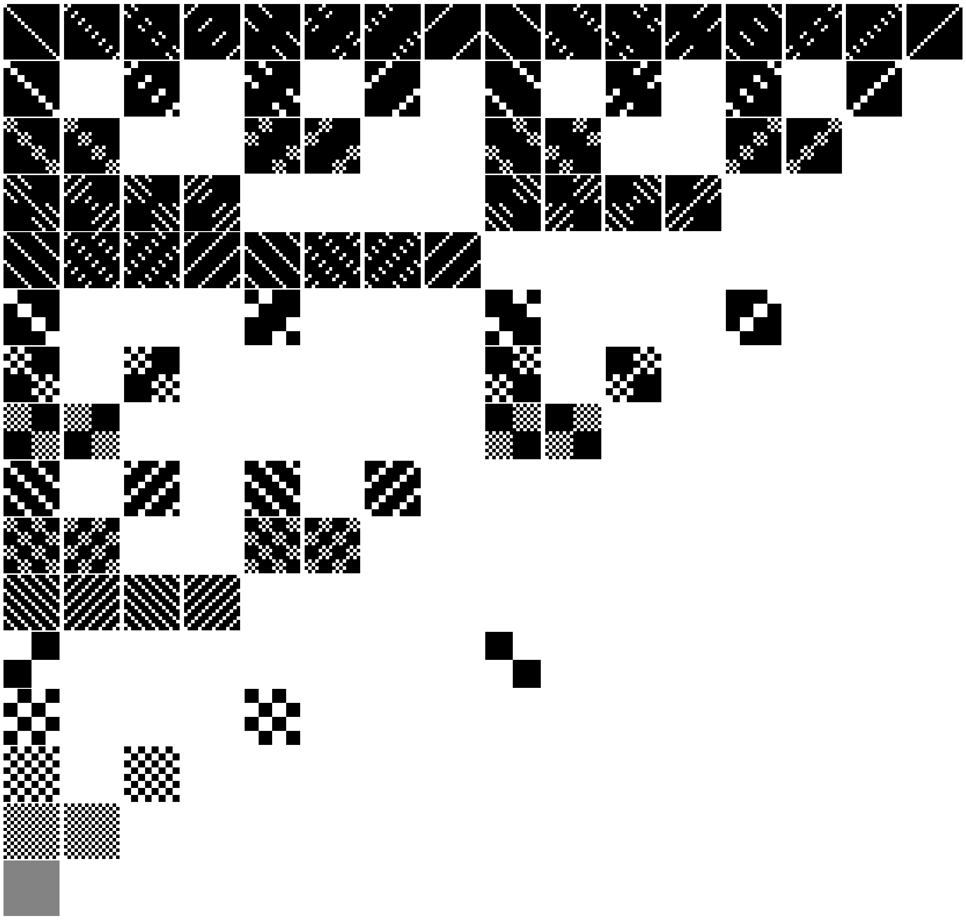

Figure 4: The kernels K_p for p the uniform distributions on faces of $\{0,1\}^4$ of dimension zero (first line), one (the next four lines; one line for each possible edge orientation), two (the next six lines; one for each pair in $\{1,2,3,4\}$), three (the next four lines), and four dimensional (p is the uniform distribution on $\{0,1\}^4$). The first row of each kernel is always equal to the probability distribution p. The rows and columns of each kernel are in the lexicographical order of $\{0,1\}^4$. By Proposition 4, all these kernels are contained in the family $\mathcal{K}_{4,4}$.